Discovering Organizational Correlations from Twitter

Jingyuan Zhang*, Xiaoxiao Shi*, Xiangnan Kong†, Hong-Han Shuai‡ and Philip S. Yu*
*Department of Computer Science, University of Illinois at Chicago, IL, USA;
jzhan8@uic.edu, xshi9@uic.edu, psyu@cs.uic.edu
†Department of Computer Science, Worcester Polytechnic Institute, MA, USA; xkong@wpi.edu
‡Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan; d99942020@ntu.edu.tw

***Abstract*—Organizational relationships are usually very complex in real life. It is difficult or impossible to directly measure such correlations among different organizations, because important information is usually not publicly available (e.g., the correlations of terrorist organizations). Nowadays, an increasing amount of organizational information can be posted online by individuals and spread instantly through Twitter. Such information can be crucial for detecting organizational correlations. In this paper, we study the problem of discovering correlations among organizations from Twitter. Mining organizational correlations is a very challenging task due to the following reasons: a) Data in Twitter occurs as large volumes of mixed information. The most relevant information about organizations is often buried. Thus, the organizational correlations can be scattered in multiple places, represented by different forms; b) Making use of information from Twitter collectively and judiciously is difficult because of the multiple representations of organizational correlations that are extracted. In order to address these issues, we propose multi-CG (multiple Correlation Graphs based model), an unsupervised framework that can learn a consensus of correlations among organizations based on multiple representations extracted from Twitter, which is more accurate and robust than correlations based on a single representation. Empirical study shows that the consensus graph extracted from Twitter can capture the organizational correlations effectively.**

I. Introduction

The activities of organizations are usually very complex in real life, consisting of abundant interactions with other organizations. For example, Microsoft collaborates with Nokia to release the Windows Phone and competes with Google and Apple on the smartphone market. Discovering the complex correlations among different organizations is very important to many real-world applications, such as corporate fraud detection and organizational activity analysis. One example is the Enron Scandal where the unusual relationships among Enron, its offshore subsidiaries, and related companies were initially ignored by the public, but eventually caused the largest corporate bankruptcy in U.S. history. Another example is the media monopoly where the opinions of different media are controlled to be in a unanimous agreement by some evil backstage manipulators. If we can automatically discover the correlations among different organizations, we can use them to understand hidden connections among organizations. However, directly measuring the correlations among different organizations is difficult or impossible, because important information about organizational relationships is not publicly available. For instance, information about terrorist organizations is usually kept secret, making it very hard to discover correlations among them.

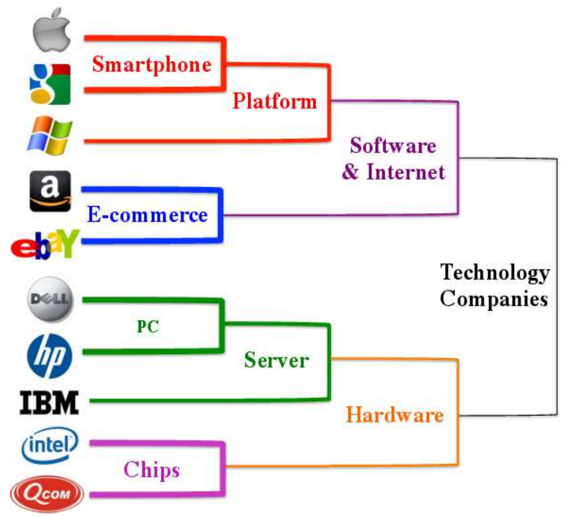


Figure 1. Organizational correlations discovered by mining Twitter data. The thickness of the lines indicates the strength of correlation discovered from Twitter posts. The potential real life meaning of each correlation pair is annotated.

With the development of Twitter, an increasing amount of organizational information can be posted online by individuals and spread instantly. Such information can be crucial for detecting organizational correlations. Figure 1 illustrates the correlations between technology companies that we detected by mining Twitter data. The thickness of the lines shows the degree of correlation we discovered from Twitter posts. We also annotated the potential real life meaning of each correlation pair. As a result, real-world correlations among a set of organizations can potentially be discovered by mining Twitter data without requiring additional information sources.

Many previous works have been proposed on mining Twitter data [3], [4], [8], [17], [19], [25], [26]. In these works, Twitter is used as a free "sensor network" to "sense" certain signals, which are expensive or hard to measure in

real life. For example, researchers can use Twitter to detect real-time events (such as earthquakes and hurricanes) around the world [19], analyze the influence of public mood on stock market [3], [4], investigate the credibility of topics spreading through Twitter [8], etc. However, all of these works mainly focus on discovering independent signals.

In this paper, we use Twitter to discover the correlations of organizations. Mining organizational correlations is a very challenging task due to the following reasons:

- Data in Twitter occurs as large volumes of mixed information. The most relevant information about organizations is often buried. Thus, the organizational correlations can be scattered in multiple places or sources under different representations. For instance, two organizations can be correlated with each other because they frequently co-appear in the same tweets. The two organizations may also be related if the changes of the number of tweets mentioning each organization are correlated. The information from a single representation inadequately reflects the real-world organizational correlations. This makes identifying the organizational correlations comprehensively from Twitter extremely challenging.
- Another challenge for discovering organizational relationships from Twitter lies in the fact that there are different types of relationships that can be extracted. How can one combine the multiple types of relationships and compute the consensus relationships among organizations? Previous works on multi-view learning [6], [9], [13] mainly focus on clustering and classification problems, which are not applicable to the problem of discovering organizational correlations. Since each representation may play a different role in reflecting the real-world relationships for different organizations, it is a nontrivial task to combine them to find the fundamental factor that forms the organizational correlations in real life.

In order to address these issues, we propose multi-CG (**multi**ple **C**orrelation **G**raphs based model), an unsupervised framework that can learn a consensus of correlations among organizations based on multiple representations extracted from Twitter. multi-CG produces correlations that are more accurate and robust than those based on a single representation. The entire process of multi-CG consists of three steps. First, we identify the important latent factors related to organizations in Twitter to comprehensively represent the organizational information. Afterward, we build multiple correlation graphs based on those latent factors. Specifically, in each of the graphs, the nodes are the organizations, and the weighted edges represent the strength of their correlation via the corresponding factor (e.g., correlations in terms of tweet volumes). Since each factor may contain a portion of the real-world organizational correlations, we leverage the concept of coordinate descent [14] to obtain an optimal correlation graph that maximizes the consensus of all latent factors.

The contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first work of discovering the correlations of organizations. To address this important issue, we propose an unsupervised multi-CG model to learn a consensus of correlations among organizations based on multiple representations extracted from Twitter.
- In order to find the optimal correlation graph that maximizes the consensus of all latent factors, we leverage the concept of coordinate descent [14] to efficiently solve the optimization problem and guarantee the convergence of the proposed multi-CG algorithm.
- We conduct experiments on 100 public companies[1] in the U.S. market and release the dataset[2] to public. The organizational correlations discovered from Twitter are verified by the correlations among stocks of companies. The experimental results demonstrate that multi-CG outperforms the baseline methods by 23% on average.

The rest of the paper is organized as follows. Section 2 defines the problem; Section 3 describes the proposed model framework and explains the algorithm for learning from multiple correlation matrices; Section 4 presents the experimental setup and the results; Section 5 discusses related work; and Section 6 concludes the paper.

II. Problem Definition

Suppose we have a correlation discovery task for n organizations from Twitter data D. The set of all organizations is denoted as $V = \{v_1, v_2, ..., v_n\}$. D consists of all the information related to the n organizations in Twitter, i.e., the set of all tweets and retweets about the organizations in V. For each organization or company, we could use certain symbols to extract the related information from Twitter, such as the hashtags. In this way, we can construct D with less noise information in Twitter. In order to extract the relationships among different organizations, we first identify m different important representations (or factors) from D, e.g., the tweet or retweet volume representation. Then based on each representation, we build a correlation graph G from D as a weighted undirected graph $G = (V, E, A)$ where

- V represents the set of organizations. For a time period of length T, we denote the representation feature as $A \subset D$. $\forall v \in V, A(v)$ denotes all the information of v under the certain representation, e.g., the daily number of relevant tweets or retweets in D.
- $E \in \mathcal{R}^{|V|\times|V|}$ is called the correlation matrix such that $E(i,j) = e$ is the correlation of an organization pair (v_i, v_j). e is calculated according to A as shown

[1]http://en.wikipedia.org/wiki/S%26P_100.

[2]https://www.dropbox.com/s/cgd2dzo1he24285/tweets_10252012_02202013.tar.gz.

Table I

NOTATION DESCRIPTIONS FOR SEVERAL TYPES OF IMPORTANT LATENT FACTORS GENERATED FROM TWITTER

Latent Factors	Graph	Matrix	Feature	Descriptions	Abbreviations
Volume	G_t	E_t	A_t	Correlations according to the time series of tweet numbers	t
	G_r	E_r	A_r	Correlations according to the time series of retweet numbers	r
Time	G_t^l	E_t^l	A_t^l	Correlations according to the time series of tweet numbers with time lag l	$t(l)$
	G_r^l	E_r^l	A_r^l	Correlations according to the time series of retweet numbers with time lag l	$r(l)$
Co-appearance	G_{ct}	E_{ct}	A_{ct}	Correlations according to the times two organizations co-appeared in the same tweet	ct
	G_{cr}	E_{cr}	A_{cr}	Correlations according to the times two organizations co-appeared in the same retweet	cr

in Section III-A. $e > 0$ if it is a positive correlation; $e < 0$ if it is a negative correlation.

The input of the task consists of n organizations and Twitter data D. We first construct m correlation graphs from Twitter to reflect the organizational relationships comprehensively. The set of constructed graphs is denoted as $\mathcal{G} = \{G_i\}_{i=1}^m$, where $G_i = \{V, E_i, A_i\}$. Correspondingly, the set of correlation matrices can be denoted as $\mathcal{E} = \{E_i \in \mathcal{R}^{n\times n}\}_{i=1}^m$. Then we aim to distill an optimal correlation matrix O from $\mathcal{E}$ to best reflect the real-world relationships of these organizations. In order to solve this problem, we need to address the following two challenges:

1. How can one capture important latent factors of organizational correlations in Twitter comprehensively?

2. How can one learn a consensus of organizational correlations based upon multiple latent factors?

For the first challenge, we present how to identify important latent factors in Twitter and how to generate the corresponding correlation graphs in Section III-A. For the second challenge, we introduce how to judiciously integrate the different factors to find the optimal consensus in Section III-B.

III. PROPOSED METHOD

In this section, we propose multi-CG (**multi**ple **C**orrelation **G**raphs based model) to address the above two challenges. It is an unsupervised framework that can capture a consensus of correlations among organizations according to multiple representations extracted from Twitter.

A. Extracting Correlation Graphs

Recall that the input of the problem consists of n organizations and the Twitter data D. The first step of the proposed model multi-CG is to identify as many latent factors as possible, which would help understanding the organizational correlations comprehensively. Here we present several types of latent factors extracted from Twitter with their corresponding correlation graphs in detail. The notations for these graphs are shown in Table I.

1) Volume Correlation Graph: For a certain organization, it can be observed that the number of times it is mentioned in Twitter usually reflects its degree of attention among people. So if two organizations correlate with each other in reality, their volume changes can also be correlated in Twitter. We capture the correlations for organization pairs according to the time series [11], [15] of tweet numbers for a period (e.g., a month). So the correlation graph $G_t = (V, E_t, A_t)$ can be constructed as follows: Let A_t denote all the information of tweets related to the n organizations in a time period of length T. $\mathbf{x}_t = (x_1, ..., x_T)$ is extracted from A_t. It denotes the series of daily number of tweets about organization $v_i \in V$ for the time length of T. $\mathbf{y}_t = (y_1, ..., y_T)$ is the series for another organization $v_j \in V$. According to $\mathbf{x}_t$ and $\mathbf{y}_t$, we can fill out all the entries of the correlation matrix E_t by measuring the correlation $E_t(i,j)$ between each organization pair v_i and v_j. To do so, we consider $\mathbf{x}_t$ and $\mathbf{y}_t$ as samples and use Pearson Correlation [23] to set the value e_t of $E_t(i,j)$ as follows:

$$\begin{aligned} E_t(i,j) &= e_t = corr(\mathbf{x}_t, \mathbf{y}_t) \\ &= \frac{1}{T-1}\sum_{k=1}^{T}(\frac{x_k - \bar{x}_t}{s_\mathbf{x}})(\frac{y_k - \bar{y}_t}{s_\mathbf{y}}) \end{aligned} \tag{1}$$

where $\bar{x}_t$ and $s_\mathbf{x}$ are the sample mean and sample standard deviation, respectively.

Since the number of retweets shows people's interest in certain events about an organization, the retweet volume changes can also represent the organizational correlations. We can consider the retweet volumes as another latent factor, and construct the correlation graph $G_r = (V, E_r, A_r)$ in a similar way.

2) Time Correlation Graph: Due to the time delays from the real-world events to people's postings in Twitter, we can add a time delay window of size l to the pairs of tweets and retweets number series so that a new type of organizational correlation graph can be generated via this factor, i.e., time correlation graph. Take the construction of $G_t^l = (V, E_t^l, A_t^l)$, the time correlation graph of tweets with time lag l as an example. $\mathbf{x}_t = (x_1, ..., x_T)$ still denotes the series of daily tweet numbers of organization $v_i \in V$ in a time length of T. In order to detect the correlation of organization pair (v_i, v_j), we set the start date as l days previous (if l is negative), or afterwards (if l is positive) for v_j. So the time series is denoted as $\mathbf{y}_t^l = (y_{l+1}, ..., y_{T+l})$. Using Equation

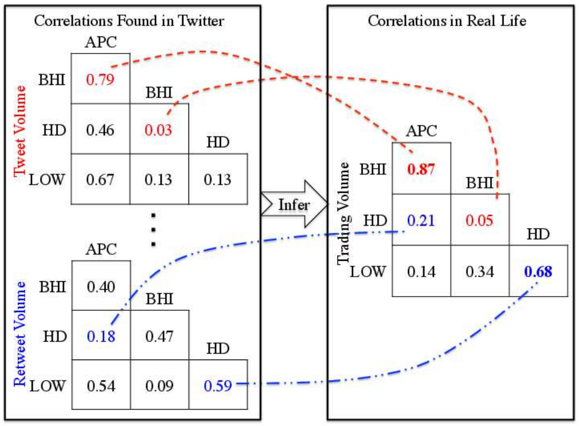


Figure 2. An example of correlations for four organizations on Oct. 25th, 2012. Each number represents the correlation value calculated under a certain factor. The red line represents the importance of tweet volume in inferring the real-world correlations for organization pairs (APC, BHI) and (BHI, HD), while the blue line represents the importance of retweet volume for another two pairs (APC, HD) and (HD, LOW).

(1) we can generate multiple correlation graphs by setting different l values for the tweet and retweet volumes.

3) Co-appearance Correlation Graph: We can also observe that two organizations are often mentioned together in Twitter if they have certain relationships in the real world. In order to capture such information, we can build a correlation graph $G_{ct} = (V, E_{ct}, A_{ct})$ according to the number of times two organizations co-appeared in the same tweet. For example, a tweet "*Maybe **Apple** should follow **Amazon** lead and make no money, that's obviously how you get a reasonable P/E multiple*." indicates that there is some kind of relationship between Apple and Amazon. Thus, let A_{ct} denote all the information of tweets related to the n organizations in a time period of length T. Let $S_i \subset A_{ct}$ denote the set of tweets that talk about v_i in a period of time (e.g., a month), and let S_j denote the set for v_j. The correlation matrix E_{ct} can be constructed in a totally different way by setting

$$E_{ct}(i,j) = e_{ct} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{2}$$

where the set $S_i \cap S_j$ corresponds to the tweets mentioning v_i and v_j simultaneously, while the set $S_i \cup S_j$ is about tweets mentioning v_i or v_j or both v_i and v_j. Similarly, a correlation graph $G_{cr} = (V, E_{cr}, A_{cr})$ can be constructed according to the number of times two organizations co-appeared in the same retweet. It is important to note that the above kinds of latent factors are just a subset of the rich Twitter information. We can identify other factors and generate the correlation graphs using the same methodology. Suppose we identify m different factors from the Twitter data D for the n organizations. We next introduce how to make good use of these different factors to find the optimal consensus, to tackle the second challenge as mentioned before.

B. multi-CG Model

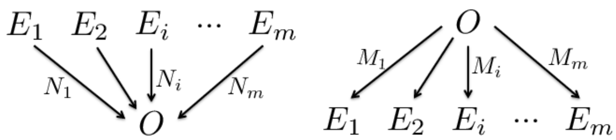
In this section, we study how to find the fundamental factors that form the structure of different correlation graphs. Recall that we have n organizations, where each contains m factors constructed from the Twitter data. Furthermore, we constructed a graph for each of the factors by capturing their underlying organizational correlations. The set of constructed graphs is $\mathcal{G} = \{G_i\}_{i=1}^m$, which reflects the organizational correlations for different factors. Correspondingly, the set of correlation matrices is $\mathcal{E} = \{E_i \in \mathcal{R}^{n \times n}\}_{i=1}^m$. Since each single factor cannot reflect the real-world organizational correlations well, it is important to find the fundamental factor that forms the structure of different correlation graphs. Hence, our objective is to learn an optimal matrix $O \in \mathcal{R}^{n \times n}$ from multiple matrices $\mathcal{E}$.

Intuitively, the optimal matrix should be a consensus of all the factors that is consistent with most factors as much as possible. More specifically, we seek for the optimal matrix that is the closest to all the matrices under certain distance measure. Given a set of correlation matrices $\mathcal{E} = \{E_i \in \mathcal{R}^{n \times n}\}_{i=1}^m$, a set of non-negative weights $\{w_i \in \mathcal{R}_+\}_{i=1}^m$ and a distance function d, the optimal correlation matrix $O \in \mathcal{R}^{n \times n}$ can be estimated by the minimization,

$$\min_{O \in \mathcal{R}^{n \times n}, E_i \in \mathcal{R}^{n \times n}} \sum_{i=1}^{m} w_i d(E_i, O). \tag{3}$$

Using Euclidean distance for the distance function d, we can get $O = \frac{1}{m}(\sum_{i=1}^m w_i E_i)$. It means the optimal correlation value for any organization pair (p, q) is the weighted average of the corresponding correlation values from all the factors. However, this model is oversimplified, which assumes that, for any organization pair (p, q), each factor plays the same role in estimating the optimal matrix O, since the weight w_i is a constant for its corresponding factor. In Twitter, this is not often the case. Figure 2 gives an example of correlations for four organizations on Oct. 25th, 2012. APC (Anadarko Petroleum Corporation) and BHI (Baker Hughes Incorporated) are both oil companies, and they have a high correlation (0.87) on trading volume of the stock market. So does another two home improvement stores HD (Home Depot) and LOW (Lowe's). We only show two latent factors, i.e., tweet and retweet volume, in the left sub-figure. It can be observed that the tweet factor is more important than the retweet factor for organization pairs (APC, BHI) and (BHI, HD). While for another two pairs (APC, HD) and (HD, LOW), the retweet factor is relevant to the ground truth. It is thus ineffective to select a set of universal weights $\{w_i \in \mathcal{R}_+\}_{i=1}^m$ for all organization pairs.

To address the above issue, we consider the organization pairs under each factor discriminatingly. Our goal is to map all the factors $\mathcal{E} = \{E_i \in \mathcal{R}^{n \times n}\}_{i=1}^m$ to a common matrix O to capture the commonality among these factors. Meanwhile we also have to ensure the distillation from the original matrix is minimal. Such distillation can be measured by mapping O to the original matrices.



(a) Distillation (b) Restoration

Figure 3. Two perspectives of the framework

Hence, the objective can be shown in either Figure 3 (a) or Figure 3 (b), which is equivalent to each other. In Figure 3 (a), N_i is defined as a distillation matrix for a factor and $E_i N_i$ extracts the commonality between E_i and O. So we expect $E_i N_i$ and O can be as close as possible. On the other hand, as shown in Figure 3 (b), we can define M_i as a restoration (or reconstruction) matrix for a factor and OM_i represents the reconstruction of E_i. So we expect OM_i and E_i can be as close as possible.

With the concept of distillation matrix and restoration matrix, the objective function can be formulated as either

$$\min_{O\in\mathcal{R}^{n\times n}, N_i\in\mathcal{R}^{n\times n}} \sum_{i=1}^{m} d(E_i N_i, O) \tag{4}$$

or

$$\min_{O\in\mathcal{R}^{n\times n}, M_i\in\mathcal{R}^{n\times n}} \sum_{i=1}^{m} d(E_i, OM_i). \tag{5}$$

The objective functions in (4) and (5) optimize both the correlation matrix O and the distillation (restoration) matrices N_i (M_i). However, (4) has a serious problem: there always exists a trivial global optimal solution $O^* = \mathbf{0}_{n\times n}$ and $N_i^* = \mathbf{0}_{n\times n}$. In order to avoid this problem, we use the objective function in (5) to formally define the framework of multi-CG model.

With the Euclidean distance for the distance function d, the model is reduced to the following optimization,

$$\begin{aligned} &\min_{O,M_i} \sum_{i=1}^{m} \|E_i - OM_i\|_F^2 \\ &s.t. \quad \|O\|_F^2 \le 1 \end{aligned} \tag{6}$$

where $\|\bullet\|_F$ denotes Frobenius norm such that $\|X\|_F = \sqrt{\sum_{ij} x_{ij}^2}$. We use $\|O\|_F^2 \le 1$ to constrain the scale of O.

Since Frobenius norm is a separable distance function, by letting $E = [E_1, ..., E_m]$ and $M = [M_1, ..., M_m]$ we can reformulate (6) as follows.

$$\begin{aligned} &\min_{O,M} \sum_{i=1}^{m} \|E - OM\|_F^2 \\ &s.t. \quad \|O\|_F^2 \le 1. \end{aligned} \tag{7}$$

We transform (7) to an unconstrained optimization problem by adding a smoothing parameter α as in (8).

$$\min_{O,M} \sum_{i=1}^{m} \|E - OM\|_F^2 + \alpha\|O\|_F^2 \tag{8}$$

where α is a positive constant. Let $f(O, M)$ denote the objective function in (8). It can be rewritten as follows:

$$\begin{aligned} f(O,M) &= tr((E-OM)^\top(E-OM)) + \alpha\ tr(O^\top O) \\ &= tr(E^\top E) - 2tr(M^\top O^\top E) \\ &\quad + tr(M^\top O^\top OM) + \alpha\ tr(O^\top O) \end{aligned} \tag{9}$$

where $tr(X)$ is the trace of a matrix X.

In order to efficiently find the optimum, we apply the coordinate descent method [14] to iteratively update O and M until it converges. Taking the partial derivative of Equation (9) with respect to M and setting it to 0, we obtain

$$O^\top OM = O^\top E. \tag{10}$$

This gives us the update rule for M with a given O. Then by taking partial derivative of $f(O, M)$ with respective to O, we obtain the following

$$O(MM^\top + \alpha I) = EM^\top. \tag{11}$$

With Equations (10) and (11), we initialize O by an identity matrix I and iteratively update M and O. Algorithm 1 summarizes the proposed multiple correlation graphs learning framework. As can be observed, M is first updated assuming O is fixed at every iteration, and then O is sequentially updated based on M. The algorithm for multi-CG terminates when it converges to a certain value. Following the way to prove the convergence of the coordinate descent method in [14], we can guarantee the convergence of the proposed multi-CG. The setting of the parameter α will be studied in the end of Section 4.

Algorithm 1 multi-CG

Input: A set of organizations $V = \{v_1, ..., v_n\}$ and the Twitter data D

Output: An optimal matrix O and a set of restoration matrices $M = [M_1, ..., M_m]$

```
Identify m important latent factors from D
Build m correlations matrices E = [E_1, ..., E_m] corresponding to the correlations graphs G = [G_1, ..., G_m]
Initialize O to an identity matrix I
while NOT converged do
    Update M using Equation (10)
    Update O using Equation (11)
end while
```

IV. Experiments

In this section, we conduct experiments on 100 public companies in the U.S. market, and use the correlations among their stocks to verify the discovered correlations by multi-CG.

Table II
DATA DESCRIPTIONS

Start Date	End Date	#Tweets	#words
Oct. 25, 2012	Feb. 20, 2013	757,929	23,016,807

A. Dataset Preprocessing

We use Twitter API to collect the public tweets talking about the 100 companies in the U.S. market from October 25, 2012. The cashtag in Twitter (e.g., $AAPL for Apple Inc) is used to extract the related information of each company in order to reduce the noise. All these tweets with their related information are denoted as the Twitter dataset D as mentioned in Section II. The statistics of the dataset are summarized in Table II.

In order to further filter out the noise information, we apply a smoothing method after we obtain the time series of tweet and retweet volumes for each organization. The simple moving average (SMA)[3] could help smooth out the short-term fluctuations in the time series so that the real trends could be better reflected for a longer time period. SMA is the unweighted mean of the previous ℓ datum points in a time series, and it is often applied to reduce random noise. For a time series $\mathbf{x}_t = (x_1, ..., x_T)$ with the daily tweet numbers of organization v for T days, the k day simple moving average for day d is computed by:

$$SMA(d) = \frac{\sum_{i=1}^{\ell} x_{(d-i)+1}}{\ell}, \ell \leq d \tag{12}$$

The value ℓ depends on the periodic fluctuation of the data movement, such as short, intermediate, or long term. In the experiments, we set $\ell = 10$.

B. Experiment Setup

Recall that multi-CG identifies several latent factors from Twitter and builds their corresponding correlation graphs. In the experiments, we generate the different correlation graphs from Twitter as described in Table I. For the time correlation graphs, we set the time lag $l \in \{-2, -1, 1, 2\}$. The output of the model is an optimal correlation matrix O from multiple different latent factors. In order to evaluate the performance of the algorithm, we have to generate the correlation matrix observed from the real-world perspective. The correlations among their stocks are helpful to verify discovered correlations by multi-CG, since the changes on stock price, trading volume, volatility, etc. show the historical development of a company in the real world, which should be captured in Twitter. Hence, we download the daily stock market data from Yahoo! Finance[4] for the 100 companies from October 25th, 2012 to February 20th, 2013, and use the correlations between companies in terms of the daily trading volumes, closing prices and historical volatilities.

[3] http://en.wikipedia.org/wiki/Moving_average#Simple_moving_average
[4] http://finance.yahoo.com/

- **Trading Volume Correlation (TV)** Trading volume is an important indicator about the market's liquidity. Higher volume means higher liquidity. We generate the volume correlation matrix in a similar way when we deal with the tweet series correlation. We denote the correlation matrix of the trading volume as C_{tv}.
- **Closing Price Correlation (CP)** Closing price generally refers to the last price at which a stock trades during a regular trading session. For each time series of a stock, we use the logarithm of returns (log returns) of closing price to calculate the correlation value. The log return r^i_{log} at time i equals to $r^i_{log} = ln(p_i/p_{i-1})$, where p_i is the closing price at time i. The correlation matrix of the closing price is denoted as C_{cp}.
- **Historical Volatility Correlation (HV)** Historical volatility measures the fluctuation of stock price during a given time period. In order to compute the volatility correlation, we first calculate the daily log returns of closing prices. Then we measure the standard deviation (STD) of this return for the last 21 days (an average trading month). We use the STD values to calculate the volatility correlation matrix C_{hv}.

Based on the definition of correlation matrices from stock markets, we set up the evaluation criterion using discounted cumulative gain (DCG) [10] to compare the ranking quality between the optimal matrix O learned from Twitter and the stock market correlation matrix $C \in \{C_{tv}, C_{cp}, C_{hv}\}$ from Yahoo! Finance. DCG is a popular measure in information retrieval tasks, and it focuses on the correctness of highly relevant entities. In order to calculate DCG, we need a two-step processing:

Step 1: Rank each row of O and C for each company according to the absolute value of correlation in a decreasing order. Then map the ordered values to the corresponding company names for each row. We denote them as L^O and L^C, respectively. Thus, the ith company in L^O has a list L^O_i, denoting the ordered companies that correlated with it from the strongest to weakest. It can be noticed that L^C gives the right (ideal) ranking for each company.

Step 2: Provide the relevance grades for the top k ranks of each row in L^O according to the right orders in L^C as in [10]. We focus on the top k ranks because we are more interested in the highly relevant company correlations than the whole rank lists. Then we calculate the DCG value for each company and define the average value of top k as the marginal gain on DCG. The formula is $avgDCG_k(O, C) = \frac{1}{n}\sum_{i=1}^{n} DCG_k(L^O_i, L^C_i)$, where $DCG_k(L^O_i, L^C_i)$ is the DCG value for ranking list L^O_i of the ith company, L^C_i is the right orders and n is the number of companies.

So far as we know, no algorithm is available in the literature for the correlation of organization pairs. Thus, we compare the performance of multi-CG with the following baselines:

Table III

OVERALL PERFORMANCES "$avgDCG$" FOR SEVERAL DIFFERENT TIME PERIODS. "↑" INDICATES THE LARGER THE VALUE THE BETTER THE PERFORMANCE. THE BOLD NUMBER INDICATES THE BEST PERFORMANCE. "∗" INDICATES THE BEST PERFORMANCE OF MULTI-CG AMONG THE THREE INDICATORS. THE *red italic* NUMBER INDICATES THE BEST PERFORMANCE OF ALL THE LATENT FACTORS FROM TWITTER FOR THE INDICATOR OF TRADING VOLUME. THE *blue italic* NUMBER INDICATES THE BEST PERFORMANCE OF ALL THE LATENT FACTORS FROM TWITTER FOR THE PRICE RELATED INDICATORS.

(a) $avgDCG$ performances for the time period from 10/25/2012 to 11/25/2012

		Single Graph												Consensus Graph	
Indicator	Top k	SC$_t$	SC$_r$	SC$_{t(-2)}$	SC$_{t(-1)}$	SC$_{t(+1)}$	SC$_{t(+2)}$	SC$_{r(-2)}$	SC$_{r(-1)}$	SC$_{r(+1)}$	SC$_{r(+2)}$	SC$_{ct}$	SC$_{cr}$	SA	multi-CG
Trading Volume ↑	10	*4.51*	3.96	3.69	*4.51*	4.21	3.05	3.56	4.27	4.27	3.87	3.56	3.08	4.26	**4.87**∗
	20	*22.06*	20.15	20.89	21.20	21.25	18.63	17.44	20.09	20.33	18.46	18.02	16.78	20.69	**22.42**∗
	30	53.84	52.76	*54.16*	53.32	53.09	51.72	47.65	52.10	52.92	50.21	48.60	47.84	52.78	**57.08**∗
	40	102.75	106.91	106.76	*110.68*	109.42	104.81	99.10	104.23	106.51	102.44	103.27	101.36	105.84	**112.73**∗
	50	*194.45*	187.10	185.04	190.30	192.25	183.60	177.94	182.28	185.90	181.77	182.44	178.45	185.49	**197.54**∗
Closing Price ↑	10	3.20	3.18	2.43	2.81	2.80	2.35	3.20	2.17	2.07	3.22	*4.41*	4.33	2.65	**4.55**
	20	17.49	16.67	15.58	16.30	16.27	15.88	17.64	15.47	15.00	19.08	20.12	*20.54*	16.22	**22.12**
	30	48.41	46.74	45.66	46.19	45.89	46.08	48.80	46.69	44.92	51.37	51.05	*51.59*	45.67	**56.45**
	40	101.86	97.95	97.20	98.59	97.89	97.74	100.75	98.92	96.86	103.66	103.96	*104.79*	97.35	**110.71**
	50	180.62	175.29	174.33	175.62	175.95	175.52	179.02	175.97	174.92	182.78	190.10	*190.78*	175.24	**193.67**
Historical Volatility ↑	10	3.11	3.05	3.08	3.05	2.85	2.65	2.62	2.74	2.78	2.57	*3.14*	3.05	2.86	**3.81**
	20	17.52	17.59	17.90	18.11	16.14	17.09	14.72	15.11	15.90	15.38	18.13	*18.29*	16.32	**19.08**
	30	48.54	47.38	50.24	49.90	46.68	48.80	43.68	44.80	45.95	45.42	50.45	*50.58*	46.68	**51.36**
	40	101.60	99.14	100.00	100.66	99.15	100.23	93.84	98.15	99.83	95.14	*100.76*	100.36	99.35	**103.62**
	50	180.08	179.07	180.45	180.12	178.03	180.89	170.46	176.70	179.45	172.22	181.03	*181.89*	178.95	**185.49**

(b) $avgDCG$ performances for the time period from 11/01/2012 to 12/01/2012

		Single Graph												Consensus Graph	
Indicator	Top k	SC$_t$	SC$_r$	SC$_{t(-2)}$	SC$_{t(-1)}$	SC$_{t(+1)}$	SC$_{t(+2)}$	SC$_{r(-2)}$	SC$_{r(-1)}$	SC$_{r(+1)}$	SC$_{r(+2)}$	SC$_{ct}$	SC$_{cr}$	SA	multi-CG
Trading Volume ↑	10	3.93	*4.81*	3.38	3.71	3.92	3.45	3.69	3.16	3.46	3.84	4.08	4.18	4.38	**4.89**∗
	20	20.40	19.38	18.32	*22.30*	20.58	19.02	18.75	17.64	18.04	19.33	22.04	20.80	21.53	**22.59**∗
	30	53.14	50.80	50.30	52.09	*53.82*	51.40	49.67	47.22	47.37	50.52	53.74	51.75	54.15	**58.56**∗
	40	*108.99*	103.85	104.38	106.94	107.96	106.65	100.62	97.89	98.17	101.63	108.86	101.28	108.30	**115.55**∗
	50	187.16	182.10	183.04	185.15	187.08	*187.18*	178.83	173.72	174.19	179.25	185.97	174.47	187.43	**200.07**∗
Closing Price ↑	10	2.98	2.53	2.36	2.97	2.81	2.84	2.60	2.80	2.82	3.03	*4.12*	4.02	2.94	**4.56**
	20	16.01	16.40	14.90	15.82	15.37	16.02	15.67	16.00	16.23	16.90	*20.80*	20.53	15.13	**22.53**
	30	45.51	46.30	44.33	44.46	43.94	45.89	44.96	45.09	46.28	47.27	50.78	*52.99*	42.80	**55.57**
	40	95.66	96.60	94.79	94.35	93.24	96.19	96.03	93.04	95.77	99.24	103.72	*106.31*	96.02	**111.35**
	50	171.31	172.60	170.39	169.55	167.96	170.34	173.20	168.58	171.76	177.67	*189.40*	182.81	169.27	**195.04**
Historical Volatility ↑	10	3.02	2.85	2.84	2.99	3.08	2.95	3.12	2.38	2.53	2.89	3.02	*3.28*	2.95	**3.55**
	20	17.74	15.82	17.42	17.90	17.85	17.40	17.16	16.25	16.68	17.88	*18.28*	18.01	17.59	**19.88**
	30	50.69	46.38	50.22	51.26	51.19	50.80	47.87	47.28	47.72	49.82	*51.85*	51.32	50.45	**52.98**
	40	104.93	99.34	104.07	104.87	104.94	104.20	101.15	98.57	98.56	104.05	*106.03*	105.01	103.91	**108.93**
	50	186.00	179.96	185.57	185.36	186.21	*187.04*	180.26	176.72	176.35	183.28	185.60	185.43	184.84	**189.36**

(c) $avgDCG$ performances for the top 10 correlations of another three time periods

		Single Graph												Consensus Graph	
Indicator	Time Period	SC$_t$	SC$_r$	SC$_{t(-2)}$	SC$_{t(-1)}$	SC$_{t(+1)}$	SC$_{t(+2)}$	SC$_{r(-2)}$	SC$_{r(-1)}$	SC$_{r(+1)}$	SC$_{r(+2)}$	SC$_{ct}$	SC$_{cr}$	SA	multi-CG
Trading Volume ↑	11/08	*4.42*	4.19	2.95	2.87	3.19	2.85	2.43	2.94	3.29	2.77	3.49	3.08	3.73	**4.73**∗
	11/15	*4.11*	3.18	2.68	*4.11*	*4.11*	2.23	2.16	3.60	3.52	2.26	3.40	3.13	3.26	**4.47**∗
	11/23	3.48	3.53	2.68	3.44	3.53	2.65	2.63	3.50	*3.57*	2.91	3.28	3.21	3.16	**3.90**∗
Closing Price ↑	11/08	3.12	3.94	3.04	2.82	2.61	2.70	2.87	2.34	2.32	3.25	*4.35*	4.28	3.22	**4.64**
	11/15	3.19	3.38	1.82	2.48	2.50	2.83	2.87	3.15	2.97	2.85	*3.87*	3.34	2.70	**4.04**
	11/23	3.43	3.36	2.49	2.83	3.17	2.20	2.82	3.31	3.19	2.82	*3.47*	*3.47*	2.26	**3.83**
Historical Volatility ↑	11/08	2.60	3.25	2.31	2.54	2.75	2.83	2.64	2.25	2.29	2.68	*3.70*	3.55	2.65	**3.91**
	11/15	2.94	2.63	2.68	2.97	3.16	2.82	2.18	3.23	3.34	2.35	3.31	*3.39*	2.69	**3.50**
	11/23	2.68	2.93	2.32	2.68	2.99	2.96	2.73	3.29	3.19	2.46	3.04	*3.39*	2.83	**3.51**

- The **S**imple **A**veraging model as in Equation (3) (abbreviated as "SA").
- The **S**ingle **C**orrelation matrix corresponding to latent factor f, including t, r, $t(l)$, $r(l)$, ct, and cr as defined in Table I (abbreviated as "SC$_f$"). For example, $SC_{t(-1)}$ means the time correlation matrix of the tweet volume with time lag -1.

All codes were implemented in Java, and all experiments were performed on a PC running OS X with 2.90 GHz Intel Core i7 PC and 8 GB memory.

C. Result Analysis

In this section, we evaluate the performance of the proposed multi-CG model on inferring the real-world organizational correlations.

For different time periods of a month from October 25th, 2012 to February 20th, 2013, we generate the different correlation graphs from Twitter as in Table I. For the time correlation graphs, we set the time lag $i \in \{-2, -1, 1, 2\}$. Then we learn the consensus graph with the unsupervised framework multi-CG and compute the $avgDCG$ values according to the three types of stock-level correlations $C \in \{C_{tv}, C_{cp}, C_{hv}\}$. Table III shows the performances of the proposed multi-CG for different time periods compared with the baselines for different top k. Due to space limit, we only show the top 10 performances in Table III (c) for the time periods. It can be observed that:

- multi-CG can outperform the baselines for different top k values and for different types of stock-level correlations, i.e., trading volume, closing price and historical volatility (highlighted in **bold**). The performance of SA is much poorer compared with the performance of the best single factor. The reason is that, although each latent factor reflects certain organizational correlations, it also contains substantial irrelevant information. Simply combining them cannot reduce such irrelevant information.
- Twitter information best reflects the correlations for trading volume among the three indicators. It can achieve the best performance (highlighted in $*$). For the other two indicators, Twitter reveals the closing price correlations much better than the historical volatility ones. Take the top 10 performances for time period of 10/25/2012 to 11/25/2012 as an example. multi-CG can achieve 4.87 and 4.55 for trading volume and closing price correlations, respectively. But the performance is only 3.81 for historical volatility.
- Among all the single latent factors extracted from Twitter, the tweet volume captures the stock trading volume correlations more accurately than the other ones (highlighted in *red italic*). Besides, the time factors in Twitter also reflect the trading volume correlations very well. For example, SC$_t$ and SC$_{t(-1)}$ both have an $avgDCG$ value of 4.51 for the correlations of trading volume between 10/25/2012 and 11/25/2012. It is much higher than the performance of SC$_{cr}$, the retweet co-appearance factor (3.08).
- The latent factors of tweet and retweet co-appearance have an important influence on price related correlations in reality, i.e., the closing price and historical volatility (highlighted in *blue italic*). For example, SC$_{ct}$ and SC$_{cr}$ achieve 4.41 and 4.33 for the correlations of closing price between 10/25/2012 and 11/25/2012. It illustrates that two organizations' stock price or volatility correlation can be well detected in Twitter if there are many tweets or retweets talking about the two organizations simultaneously.

D. Case Study

Now we present a case study to show the effectiveness of the unsupervised multi-CG model for the real-world organizational correlations. We focus on several companies that are more related to the consumer sentiments, like LOW (a home-improvement retailer), HNZ (a food company), etc. These companies are selected according to the industries they belong to, and Table IV gives the company names, their abbreviations and the industry information they belong to. Since the Twitter information reflects the trading volume correlations in real life more accurately, we run multi-CG on these companies and show the performance according to the top 4 correlations of trading volume.

Take the company APC as an example. We demonstrate the top 4 companies discovered in Twitter by different methods for different time periods. Table V shows the performance we learned from Twitter and the stock-level results according to trading volume. Among all the single latent factors in Twitter, we present the performance of tweet volume factor, because it captures the trading volume correlations more accurately than other ones (as shown in Table III). We can observe that Twitter reflects the trading volume correlations very well in each time period. What's more, the correlation changes are captured accurately for different time intervals. For example, APC has very strong correlations with F, APA, DVN and HAL from 10/25/2012 to 11/25/2012, but from 11/01/2012, APC is no longer strongly correlated with F. multi-CG detects such changes and discoveries the strong correlation between APC and SLB after 11/01/2012, which is consistent with the real-world trading volume correlations. However, the comparative methods fail in doing so. Therefore, the proposed multi-CG model is more robust in inferring organizational correlations for different time periods.

Another thing observed from Table V is that companies in the same industry category often have high correlations. APA, APC, DVN, HAL, NOV and SLB are all related to the Oil and Gas Industry. multi-CG can learn well from this kind of knowledge and reveal the relationships which fit our intuition in real life.

Table IV
INFORMATION OF SEVERAL COMPANIES

Name	Abbreviation	Industry
Allstate Corporation	ALL	Property & Casualty Insurance
Apache Corporation Anadarko Petroleum Corporation Devon Energy Corporation	APA APC DVN	Independent Oil & Gas
Baker Hughes Incorporated Halliburton Company National Oilwell Varco Schlumberger Limited	BHI HAL NOV SLB	Oil & Gas Equipment & Services
Cisco Systems	CSCO	Networking & Communication Devices
Exelon Corporation	EXC	Diversified Utilities
Ford Motor	F	Auto Manufacturers - Major
Home Depot Lowe's	HD LOW	Home Improvement Stores
H. J. Heinz Company	HNZ	Food - Major Diversified
Metlife Incorporated	MET	Life Insurance

Table V
TOP 4 CORRELATIONS OF TRADING VOLUME (TV) FOR COMPANY APC. THE COMPANY IN RED INDICATES THE REAL-WORLD TRADING VOLUME CORRELATION FOR A CERTAIN TIME PERIOD. THE COMPANY IN BOLD INDICATES ITS APPEARANCE IN THE GROUND TRUTH (I.E., THE TV COLUMN). THE COMPANY WITH "$*$" INDICATES ITS RIGHT RANK DISCOVERED FROM TWITTER.

Rank	10/25/2012∼11/25/2012				11/01/2012∼12/01/2012				11/08/2012∼12/08/2012			
	SC_t	SA	multi-CG	TV	SC_t	SA	multi-CG	TV	SC_t	SA	multi-CG	TV
1	**APA**	**DVN**	**F**$*$	F	**MET**	**MET**	**SLB**$*$	SLB	**SLB**$*$	MET	**SLB**$*$	SLB
2	**F**	BHI	**APA**$*$	APA	F	F	**DVN**$*$	DVN	F	DVN	**NOV**$*$	NOV
3	**HAL**	**F**	**DVN**$*$	DVN	**HAL**	**DVN**	**MET**$*$	MET	**NOV**	F	**ALL**$*$	ALL
4	BHI	NOV	**HAL**$*$	HAL	**DVN**	NOV	**HAL**$*$	HAL	MET	**NOV**	MET	BHI

E. Parameter Sensitivity

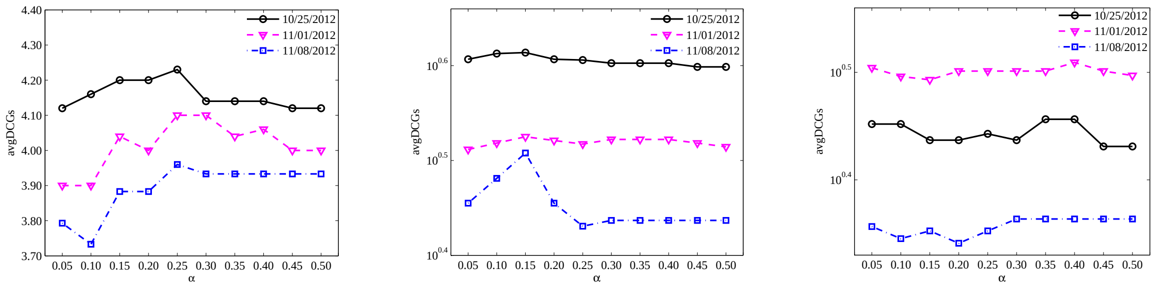
According to Equation (8), the proposed model has a smoothing parameter α. It controls the flexibility of the the objective function in Equation (7). In this section, we study the sensitivity of multi-CG to this parameter. We demonstrate the results on correlations of trading volume, closing price and historical volatility for three different time periods, respectively. In Figure 4(a) we can observe that multi-CG performs consistently on different α values for trading volume correlations. The fluctuations are in a small range, especially for monthly period of 10/25/2012. We also observe that the best performances are achieved at α = 0.25.

Figure 4(b) and Figure 4(c) show that similar effects of α can be observed on closing price and historical volatility. Though the performance varies on different α values, they have a boundary with a small range. For closing price, the highest $avgDCG$ value is achieved when α equals to 0.15, while the highest value of α is 0.4 for historical volatility. Hence, we assign 0.25 to α in all the experiments for trading volume correlations, 0.15 for closing price correlations and 0.4 for historical volatility correlations, respectively.

V. RELATED WORK

One related work to our study is analyzing social networks to better understand social behaviors and human interactions. For instance, Crandall et al. [5] studied how friendships form. Many other works also study how the information diffuses in social networks [12] and how people influence each other [1]. However, all these projects aim to discover interesting properties of human behaviors from social networks instead of the organizational relationships.

The study of correlation and causality is also related with our work. An increasing number of researches have been done in various fields, including Bioinformatics [18], [24], Economics [7], Philosophy [22], etc. For example, Granger [7] won the Nobel Prize in economic 2003 by analyzing the financial and macroeconomic data of time series. Shi et al. [20], [21] model the causalities to detect shakers in economic by analyzing the behavior of time series data via temporal correlation relationships. However, all the above methods are limited to one certain field. In our research work, the correlation knowledge we learned from time series of Twitter can be corresponded to any perspectives of organizations. Another area of related work is about the multiple heterogeneous graphs learning (e.g., [2], [6], [9], [13], [16]). It aims to learn from instances which have multiple views in different feature spaces. For instance, Long et al. [13] propose a general unsupervised framework for multiple view clustering problem to reconcile the patterns from different representations. In [6], consensus learning is



(a) Trading volumes (b) Closing prices (c) Historical volatilities

Figure 4. Sensitivity evaluation

introduced to learn from each heterogeneous feature space independently and then ensemble the results. [9] focuses on the problem of multi-task and multi-view learning. It proposes a graph-based framework to take advantage of both feature heterogeneity and task heterogeneity. Our work differs from these previous approaches. We consider the organizational correlations in Twitter from different representations to help infer the real-world relationships, which is a totally different problem.

VI. Conclusion

In this paper, we investigate the problem of detecting the organizational correlations from Twitter. In order to solve it, we propose the multi-CG model to learn a consensus of correlations based on multiple representations extracted from Twitter. After we obtain the consensus graph, we rank the correlation strengths for each company in a descending order to infer the real-world organizational correlations. Several periods of Twitter data were experimented to evaluate multi-CG. It can be clearly observed that the proposed multi-CG model outperforms the comparison algorithms.

There are several promising directions for future work. In current work, multi-CG well captures the correlations of trading volumes for organizations, so one direction of our future work is to model and predict the trading behaviors of organizations in real time. To support such an online system, an incremental algorithm is desirable for incorporating the historical trading volumes into multi-CG efficiently. In addition, since deep learning methods can learn a joint representation from multiple unlabeled text and image data, we can also utilize the deep learning framework to learn the consensus correlation graph from multiple representations of Twitter data.

Acknowledgement

This work is supported in part by NSF through grants CNS-1115234 and OISE-1129076 and US Department of Army through grant W911NF-12-1-0066.

References

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. KDD*, pages 7–15, 2008.

[2] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[3] J. Bollen and H. Mao. Twitter mood as a stock market predictor. *Computer*, pages 91–94, 2011.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[5] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. KDD*, pages 160–168, 2008.

[6] J. Gao, W. Fan, Y. Sun, and J. Han. Heterogeneous source consensus learning via decision propagation and negotiation. In *Proc. KDD*, pages 339–348, 2009.

[7] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[8] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proc. SDM*, pages 153–164, 2012.

[9] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *Proc. ICML*, pages 25–32, 2011.

[10] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[11] E. J. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[12] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[13] B. Long, P. S. Yu, and Z. Zhang. A general model for multiple view unsupervised learning. In *Proc. SDM*, pages 822–833, 2008.

[14] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[15] A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover. Exact discovery of time series motifs. In *Proc. SDM*, pages 473–484, 2009.

[16] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proc. CIKM*, pages 86–93, 2000.

[17] B. O'Connor, R. Balasubramanyan, B. Routedge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[18] C. A. Ratanamahatana, J. Lin, D. Gunopulos, E. J. Keogh, M. Vlachos, and G. Das. Mining time series data. In *Data Mining and Knowledge Discovery Handbook*, pages 1049–1077. 2010.

[19] T. Sakaki, O. M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. WWW*, pages 851–860, 2010.

[20] X. Shi, W. Fan, and P. S. Yu. Dynamic shaker detection from evolving entities. In *Proc. SDM*, pages 350–358, 2013.

[21] X. Shi, W. Fan, J. Zhang, and P. S. Yu. Discovering shakers from evolving entities via cascading graph inference. In *Proc. KDD*, pages 1001–1009, 2011.

[22] Q. Smith. Causation and the logical impossibility of a divine cause. *Philosophical Topics*, 24(1):169–191, 2010.

[23] S. M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.

[24] I. Tsamardinos. Causal data mining in bioinformatics. *European Research Consortium for Informatics and Mathematics News*, 2007(69), 2007.

[25] J. M. Xu, A. Bhargava, R. Nowak, and X. Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Proc. ECML-PKDD*, pages 644–659. 2012.

[26] H. Zhang, M. Korayem, D. J. Crandall, and G. LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *Proc. WWW*, pages 749–758, 2012.